D470–D474 *Nucleic Acids Research, 2013, Vol. 41, Database issue* *Published online 26 October 2012*
doi:10.1093/nar/gks987

Glycan fragment database: a database of PDB-based glycan 3D structures

Sunhwan Jo[1,2] and Wonpil Im[1,2,*]

[1]Department of Molecular Biosciences and [2]Center for Bioinformatics, The University of Kansas, 2030 Becker Drive, Lawrence, KS 66047, USA

Received August 9, 2012; Revised September 14, 2012; Accepted September 27, 2012

ABSTRACT

The glycan fragment database (GFDB), freely available at http://www.glycanstructure.org, is a database of the glycosidic torsion angles derived from the glycan structures in the Protein Data Bank (PDB). Analogous to protein structure, the structure of an oligosaccharide chain in a glycoprotein, referred to as a glycan, can be characterized by the torsion angles of glycosidic linkages between relatively rigid carbohydrate monomeric units. Knowledge of accessible conformations of biologically relevant glycans is essential in understanding their biological roles. The GFDB provides an intuitive glycan sequence search tool that allows the user to search complex glycan structures. After a glycan search is complete, each glycosidic torsion angle distribution is displayed in terms of the exact match and the fragment match. The exact match results are from the PDB entries that contain the glycan sequence identical to the query sequence. The fragment match results are from the entries with the glycan sequence whose substructure (fragment) or entire sequence is matched to the query sequence, such that the fragment results implicitly include the influences from the nearby carbohydrate residues. In addition, clustering analysis based on the torsion angle distribution can be performed to obtain the representative structures among the searched glycan structures.

INTRODUCTION

An oligosaccharide moiety in a glycoprotein, referred to as a glycan, comes in a diversity of sequences and structures, and specific interactions between carbohydrates and proteins are essential in many cellular events (1–3). These events require molecular recognition of specific carbohydrate structures that seems to be sensitive to small differences in carbohydrate structure. For instance, the carbohydrate structures found on a host cell receptor, which only differ by the sequence of the terminal sugar residues, are believed to be a major factor in determining the host range (e.g. swine, avian or human) of influenza viruses (4,5). In addition, glycosyl transferases and glycosidases recognize specific sequences and spatially arranged oligosaccharide chains (6,7). Thus, understanding the carbohydrate conformations will provide insight into the role of glycans in modulating many cellular events.

Analogous to protein structure, the structure of an oligosaccharide chain can be characterized by the torsion angles of glycosidic linkages between relatively rigid carbohydrate monomeric units. Considerable efforts have been already made to characterize the potential energy surface of the peptide bond conformation, and the accessible torsion angles of a peptide are well known (8–12). However, unlike proteins and peptides where the amino acid units are linearly linked together by the same peptide bonds, glycans can have branches, and each monosaccharide unit can be linked by different types of glycosidic linkages. In addition, the lack of experimentally derived atomic structures of oligosaccharides in aqueous solution makes it difficult to characterize the accessible torsion angles of a particular glycosidic linkage.

Despite the difficulties involved in crystallization, the number of glycoprotein structures deposited in the Protein Data Bank (PDB) (13) has been steadily increasing (14,15). Although far from complete, glycan structures in the PDB can be used to study the accessible glycosidic torsion angles (16–19). Unfortunately, however, extracting structural information of glycans from the PDB is not trivial because of a lack of standardized nomenclature and the way the data are presented in the PDB (3,14). Recently, Säwén *et al.* (19) analysed the accessible glycosidic torsion angles of the $\alpha(1\rightarrow 2)$ linked mannose disaccharide using the PDB glycan structures, but they had to make considerable efforts to collect and filter out erroneous PDB entries.

In this work, we present the glycan fragment database (GFDB), a database of the glycosidic torsion angles

*To whom correspondence should be addressed. Tel: +1 785 864 1993; Fax: +1 785 864 5558; Email: wonpil@ku.edu

derived from the PDB glycan structures. Carbohydrate structures in the PDB are recognized by *Glycan Reader*, an automatic sugar identification algorithm that we developed (15), instead of using the nomenclature presented in the PDB entries. The GFDB provides an intuitive glycan sequence search tool that allows the user to search complex glycan structures. After a glycan search is complete, each glycosidic torsion angle distribution of the searched glycan structures is displayed. In addition, the torsion angle distributions can be clustered to generate representative structures using the clustering analysis facility on the GFDB interface. To facilitate the conformational analysis of glycosidic linkages, the GFDB also provides various filters. In the following sections, we discuss how the glycan structural information was collected, how to search a glycan sequence and how the search results are displayed. A stepwise guide about the GFDB is also provided in http://www.glycanstructure.org/fragment-db.

GLYCAN FRAGMENT DATABASE

To recognize the PDB entries that contain carbohydrate molecules, we used *Glycan Reader* for automatic sugar identification (15). Briefly, in *Glycan Reader*, topologies of the molecules in the HETATM section of a PDB file are first generated using the atom connection information from the CONECT section. The carbohydrate candidate molecules (six-membered ring for a pyranose and five-membered ring for a furanose that are composed of only one oxygen and carbon atoms) are then identified. For each carbohydrate-like molecule, the chemical groups attached to each position of the ring and their orientations are compared with a pre-defined table to identify the correct chemical name for the carbohydrates. Glycan chains are constructed by examining the glycosidic linkages between the carbohydrate molecules that have chemical bonds between them.

Identified carbohydrate molecules are further analysed and recorded in the GFDB. First, the residue name annotated in the PDB is compared with the molecular structure. The disparity of the residue name annotation in the PDB and the actual molecular structure is common (14). Although *Glycan Reader* returns the correct carbohydrate names according to the molecular structures, such disparity could be a sign of potential error. Second, because a distorted ring geometry could mislead the interpretation of the glycosidic torsion angles, the geometry of the carbohydrate ring is calculated by virtual torsion angle definition (20) and is recorded whether it is in a chair conformation ($^1C_4$ or $^4C_1$). Finally, if the carbohydrate molecules have chemical groups (phosphate, sulphate, methyl and so forth) attached in one of the hydroxyl groups, the carbohydrates are marked as derived carbohydrates in the GFDB. The entries that belong to these cases can be excluded from the search using the filtering options, such as 'misassigned residues', 'distorted ring geometry' and 'derived carbohydrates' (see later in the text and also Figure 1).

WEB INTERFACE

The GFDB provides a glycan sequence search interface that allows the user to search complex glycan sequences (Figure 1). The search interface provides a visual guide as the user builds a complex glycan query sequence, and the interface is compatible with any modern web browser with JavaScript capability. There is a report generation facility available to generate an archived report file that contains all the raw data for a given search and 3D structures based on the clustering analysis (see later in the text); the user also can get the archived report file by email. There are several filtering functions available (Figure 1), which narrows the search results for specific needs, such as filters for only N-/O-glycosylated glycans, the resolution of the PDB entries and/or the aforementioned three structural features (misassigned residues, distorted ring geometry and derived carbohydrates).

When analysing the glycosidic torsion angles in the PDB, it is important to understand that there are redundant PDB entries from the same or similar proteins. Without removing those redundant entries, it is possible to overestimate the preference of a certain conformation for a given glycan sequence. Although redundancies in the PDB can be removed by post-processing the data obtained by the GFDB, the GFDB provides a preliminary filter option for removing such redundant protein entries for N-linked or O-linked glycan chains based on the sequence similarity of the parent protein.

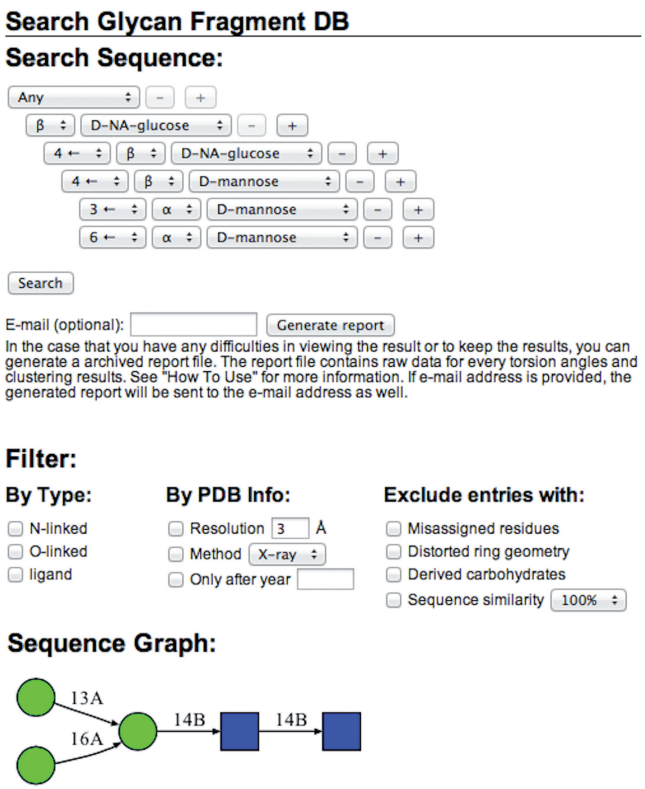

Figure 1. The GFDB search interface.

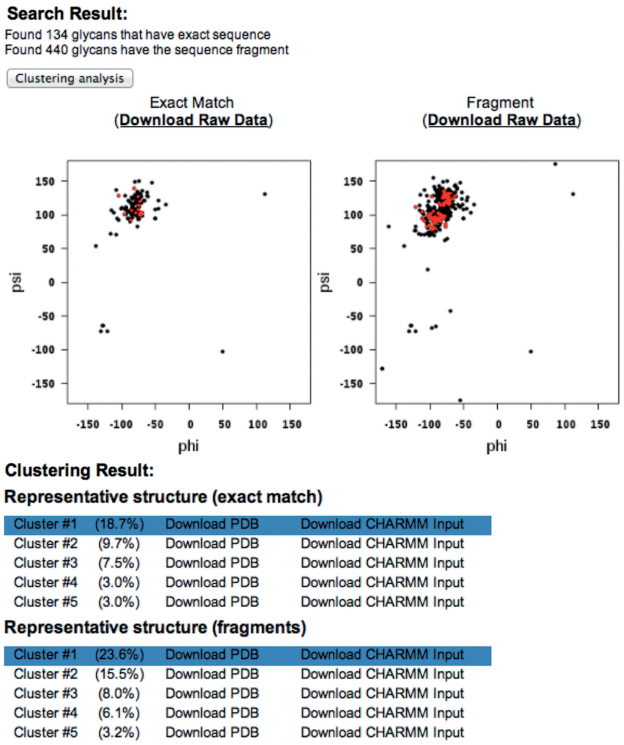

Figure 2. An example of the search result for the query sequence in Figure 1. The glycosidic torsion angle distribution of a particular glycosidic linkage can be displayed by clicking the glycosidic linkage in 'Sequence Graph' in Figure 1. The clustering analysis of the glycan chain can be performed, and the top-five representative structures can be downloaded. The glycosidic torsion angle distribution of a selected cluster is shown in red.

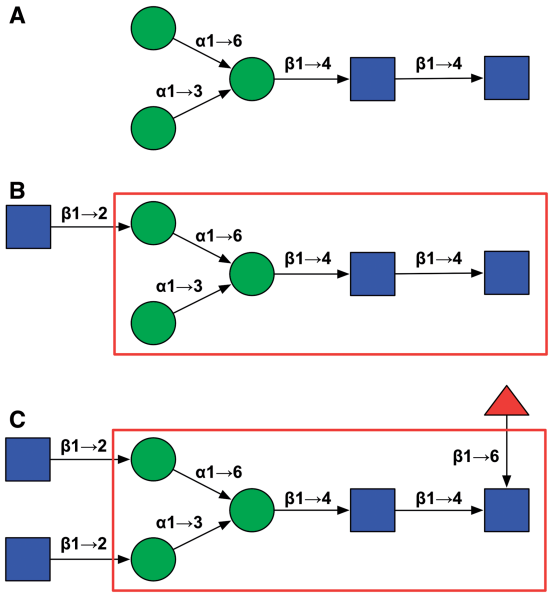

Figure 3. An example of the exact and fragment matches based on the query sequence in Figure 1. (**A**) The glycan sequence for the exact match results. (**B** and **C**) Examples of the glycan sequences for the fragment match results. The matched substructure is highlighted in the red rectangles. The sequence in (A) is also included in the fragment match results.

After a glycan search is finished, the interface shows two torsion angle distributions side by side (Figure 2), 'exact match' and 'fragment match' (Figure 3). For the exact match, the GFDB first performs a sequence search to find the PDB entries that contain the glycan sequence identical to the query sequence, and the resulting torsion angle values for each glycosidic linkage are displayed to the user. On the other hand, the fragment search performs a search against the substructures (hence, they are called fragments, Figure 3) and returns the entries having at least one substructure that matches to the query sequence. This provides more samples for the torsion angle analysis. The torsion angle values from the fragment match always contain the exact match results. However, the fragment search results may not be the same as the exact match results because part of a glycan structure can adopt a different structure when it has extra intra- and intermolecular interactions. Therefore, the fragment match results implicitly include the influences from the nearby carbohydrate residues and different protein–carbohydrate interactions, such that one can assess the flexibility of a certain glycosidic linkage in the context of larger glycan chain by comparing the exact and fragment match results.

The glycosidic torsion angle definition in the GFDB is adopted from the crystallographic definition; O_5-C_1-O_1-C'_x (ϕ), C_1-O_1-C'_x-C'_{x-1} (ψ), and O_1-C'_6-C'_5-O'_5 (ω). The torsion angle between the first residue of the N-glycan chain and the side chain of the asparagine residue is defined as O_5-C_1-N'_{D2}-C'_G (ϕ) and C_1-N'_{D2}-C'_G-C'_B (ψ). The torsion angle between the first residue of the O-glycan chain and the side chain of the serine residue is defined as O_5-C_1-O'_G-C'_B (ϕ) and C_1-O'_G-C'_B-C'_A (ψ). For threonine, O_{G1} is used instead of O_G. The atom names are based on the CHARMM topology.

CLUSTERING ANALYSIS

Statistical analysis of the torsion angle values of a particular glycosidic linkage is useful to estimate the allowable conformations of glycan chains, but it is difficult to understand what would be the representative (or most probable) structures of the given glycan sequence among the available PDB glycan structures. To provide useful insight into the 3D glycan structure, the GFDB provides an option to perform clustering analysis of the torsion angle search results and produce the top-five most clustered glycan-only structures.

The GFDB uses a simple clustering method to efficiently determine the members of each cluster. The pairwise torsion angle differences are first calculated by the following equation:

$$d_{ij} = \sqrt{\frac{\sum_k \left(\phi_i^k - \phi_j^k\right)^2 + \left(\psi_i^k - \psi_j^k\right)^2}{N}} \quad (1)$$

where N is the total number of glycosidic linkages in a glycan sequence, ϕ^k and ψ^k are the torsion angle values of the k-th glycosidic linkage, and i and j represent two glycan structures. ω torsion angle values are included only for

glycosidic linkages that have three rotatable bonds such as 1–6 linkages. After the pairwise distance matrix of the searched glycan structures is calculated, the first cluster is identified with the maximum number of neighbours within a 30° cut-off radius; the cut-off value was empirically determined. The second cluster is identified in the same manner after excluding the members that belong to the first cluster. The result of the top-five clusters and the corresponding 3D glycan structure based on the centroid of each cluster is provided to the user along with the input files to generate the centroid glycan structures using the CHARMM biomolecular simulation program (21).

CONCLUDING DISCUSSION

There are several databases that provide information on glycan structures or sequences derived from the PDB (or from other experiments). Many of these databases, such as BCSDB (22), KEGG GLYCAN (23) and Glycoconjugate Data Bank (24), store only glycan sequence information, whereas the GFDB focuses on the 3D glycan structure. GlycoMaps DB (25) and GlyTorsion (26) provide torsion angle distributions of glycosidic linkages derived from computational calculations and from the PDB, respectively. Thus, the GlyTorsion database is the only database that can be directly compared with the GFDB. While the search interface of the GlyTorsion database is restricted to only one glycosidic linkage, the GFDB can search more complex glycan sequence with various filter functions and provide the clustering analysis and the top-five clustered structures. These unique features in the GFDB allow researchers to collect complex glycan structural information easily and reliably.

As of August 2012, the GFDB contains 5360 PDB entries that contain at least one carbohydrate molecule and 20 467 glycan chains. Among those glycan chains, 11 735 (57%) are N-linked glycan chains and 788 (4%) are O-linked glycans. And the remaining 7944 (39%) exist as ligands. For the glycan structures with more than two carbohydrates, the hierarchical fragmentation identified a total of 81 370 fragment structures with 4267 unique glycan sequences; a unique glycan sequence has more than two carbohydrates and is defined by the carbohydrate sequence and the glycosidic linkages. There are 30 375 glycosidic torsion angle values available in the GFDB. By providing the straightforward search tool, the filtering functions and the clustering analysis for the representative structures, we hope that the GFDB can help conformational analysis of various oligosaccharide chains and glycosidic linkages. The database will be updated quarterly and is freely available at http://www.glycanstructure.org.

FUNDING

University of Kansas General Research Fund [2301388-003]; Kanas-COBRE NIH P20 [GM103420]; NSF [MCB-0918374]; TeraGrid/XSEDE resources [TG-MCB070009]. Funding for open access charge: NSF [MCB-0918374].

Conflict of interest statement. None declared.

REFERENCES

1. Rudd,P.M., Wormald,M.R. and Dwek,R.A. (2004) Sugar-mediated ligand–receptor interactions in the immune system. *Trends Biotech.*, **22**, 524–530.
2. Petrescu,A.-J., Wormald,M.R. and Dwek,R.A. (2006) Structural aspects of glycomes with a focus on N-glycosylation and glycoprotein folding. *Curr. Opin. Struct. Biol.*, **16**, 600–607.
3. Wormald,M.R., Petrescu,A.-J., Pao,Y., Glithero,A., Elliott,T. and Dwek,R.A. (2002) Conformational studies of oligosaccharides and glycopeptides: Complementarity of NMR, X-ray crystallography, and molecular modelling. *Chem. Rev.*, **102**, 371–386.
4. Shinya,K., Ebina,M., Yamada,S., Ono,M., Kasai,N. and Kawaoka,Y. (2006) Avian flu: influenza virus receptors in the human airway. *Nature*, **440**, 435–436.
5. Skehel,J.J. and Wiley,D.C. (2000) Receptor binding and membrane fusion in virus entry: the influenza hemagglutinin. *Annu. Rev. Biochem.*, **69**, 531–569.
6. Shah,N., Kuntz,D.A. and Rose,D.R. (2008) Golgi alpha-mannosidase II cleaves two sugars sequentially in the same catalytic site. *Proc. Natl Acad. Sci. USA*, **105**, 9570–9575.
7. Zhong,W., Kuntz,D.A., Ember,B., Singh,H., Moremen,K.W., Rose,D.R. and Boons,G.-J. (2008) Probing the substrate specificity of Golgi alpha-mannosidase II by use of synthetic oligosaccharides and a catalytic nucleophile mutant. *J. Am. Chem. Soc.*, **130**, 8975–8983.
8. Ramachandran,G.N., Ramakrishnan,C. and Sasisekharan,V. (1963) Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.*, **7**, 95–99.
9. Ramachandran,G.N. and Sasisekharan,V. (1968) Conformation of polypeptides and proteins. *Adv. Protein Chem.*, **23**, 283–438.
10. Hovmoller,S., Zhou,T. and Ohlson,T. (2002) Conformations of amino acids in proteins. *Acta Crystallogr. D Biol. Crystallogr.*, **58**, 768–776.
11. Ho,B.K., Thomas,A. and Brasseur,R. (2003) Revisiting the Ramachandran plot: hard-sphere repulsion, electrostatics, and H-bonding in the alpha-helix. *Protein Sci.*, **12**, 2508–2522.
12. Porter,L.L. and Rose,G.D. (2011) Redrawing the Ramachandran plot after inclusion of hydrogen-bonding constraints. *Proc. Natl Acad. Sci. USA*, **108**, 109–113.
13. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
14. Lütteke,T., Frank,M. and von der Lieth,C.-W. (2004) Data mining the protein data bank: automatic detection and assignment of carbohydrate structures. *Carbohydr. Res.*, **339**, 1015–1020.
15. Jo,S., Song,K.C., Desaire,H., Mackerell,A.D. Jr and Im,W. (2011) Glycan Reader: automated sugar identification and simulation preparation for carbohydrates and glycoproteins. *J. Comput. Chem.*, **32**, 3135–3141.
16. Petrescu,A.J., Petrescu,S.M., Dwek,R.A. and Wormald,M.R. (1999) A statistical analysis of N- and O-glycan linkage conformations from crystallographic data. *Glycobiology*, **9**, 343–352.
17. Petrescu,A.-J., Milac,A.-L., Petrescu,S.M., Dwek,R.A. and Wormald,M.R. (2004) Statistical analysis of the protein environment of N-glycosylation sites: implications for occupancy, structure, and folding. *Glycobiology*, **14**, 103–114.
18. Lütteke,T. (2009) Analysis and validation of carbohydrate three-dimensional structures. *Acta Crystallogr. D Biol. Crystallogr.*, **65**, 156–168.
19. Säwén,E., Massad,T., Landersjö,C., Damberg,P. and Widmalm,G.R. (2010) Population distribution of flexible molecules from maximum entropy analysis using different priors as background information: application to the Φ, Ψ-conformational space of the α-(1—>2)-linked mannose disaccharide present in N- and O-linked glycoproteins. *Org. Biomol. Chem.*, **8**, 3684–3695.
20. Rao,V.S.R. (1998) *Conformation of Carbohydrates*. Harwood Academic Publishers, Australia.
21. Brooks,B.R., Brooks,C.L., Mackerell,A.D. Jr, Nilsson,L., Petrella,R.J., Roux,B., Won,Y., Archontis,G., Bartels,C., Boresch,S. *et al.* (2009) CHARMM: the biomolecular simulation program. *J. Comput. Chem.*, **30**, 1545–1614.

22. Herget,S., Toukach,P.V., Ranzinger,R., Hull,W.E., Knirel,Y.A. and von der Lieth,C.W. (2008) Statistical analysis of the Bacterial Carbohydrate Structure Data Base (BCSDB): characteristics and diversity of bacterial carbohydrates in comparison with mammalian glycans. *BMC Struct. Biol.*, **8**, 35.
23. Hashimoto,K., Goto,S., Kawano,S., Aoki-Kinoshita,K.F., Ueda,N., Hamajima,M., Kawasaki,T. and Kanehisa,M. (2006) KEGG as a glycome informatics resource. *Glycobiology*, **16**, 63R–70R.
24. Nakahara,T., Hashimoto,R., Nakagawa,H., Monde,K., Miura,N. and Nishimura,S.-I. (2008) Glycoconjugate Data Bank: structures—an annotated glycan structure database and N-glycan primary structure verification service. *Nucleic Acids Res.*, **36**, D368–D371.
25. Frank,M., Lutteke,T. and von der Lieth,C.W. (2007) GlycoMapsDB: a database of the accessible conformational space of glycosidic linkages. *Nucleic Acids Res.*, **35**, 287–290.
26. Lütteke,T., Frank,M. and von der Lieth,C.-W. (2005) Carbohydrate Structure Suite (CSS): analysis of carbohydrate 3D structures derived from the PDB. *Nucleic Acids Res.*, **33**, D242–D246.